arXiv: math.PRxxxx.xxxxx

The Bayesian Analysis of Complex, High-Dimensional Models: Can it be CODA?

Y. Ritov[§], **P. J. Bickel**[*], **A. C. Gamst**[†], **B. J. K. Kleijn**[‡],

Department of Statistics, The Hebrew University, 91905 Jerusalem, Israel; e-mail: yaacov.ritov@gmail.com*; url:* http://pluto.mscc.huji.ac.il/∼yaacov
Department of Statistics, University of California, Berkeley, CA 94720-3860, USA; e-mail: bickel@stat.berkeley,edu*; url:* http://www.stat.berkeley.edu/∼bickel
Biostatistics and Bioinformatics, University of California, San Diego, CA 92093-0717, USA; e-mail: acgamst@math.ucsd.edu*; url:* http://biostat.ucsd.edu/acgamst.htm
Korteweg-De Vries Instituut voor Wiskunde, POSTBUS 94248, 1090 GE Amsterdam, Kamer: C4.135; The Netherlands; e-mail: B.J.K.Kleijn@uva.nl*; url:* http://home.medewerker.uva.nl/b.j.k.kleijn

Abstract: We consider the Bayesian analysis of a few complex, high-dimensional models and show that intuitive priors, which are not tailored to the fine details of the data model and the estimated parameters are going to fail in situations in which simple good frequentist estimators exit. The models we consider are, partially observed sample, the partial linear model, estimating linear and quadratic functionals of a white noise models, and estimating with stopping times. We argue that these findings do not contradict a strong version of Doob's consistency theorem which claims that the existence of a uniformly $\sqrt{n}$ consistent estimator ensures that the Bayes posterior is $\sqrt{n}$ consistent for values of the parameter with prior probability 1.

Keywords and phrases: Foundations, CODA, Bayesian inference, White noise models, Partial linear model, Stopping time, Functional estimation, Semiparametrics.

1. Introduction

We study in this paper a few examples of Bayesian procedures on complex, high-dimensional parameter spaces. Bayesian procedures can be considered from different points of view. Their closure is the set of admissible procedures, and they are known to generate asymptotic minimax procedures in regular parametric models. These and similar notions are frequentist in nature, and are not the main focus of the Bayesian paradigm.

The Bayesian procedures we consider are those that adhere to the following paradigm. The prior distribution is announced *prior* to observing the data. If this is viewed as too unrealistic, we at least restrict to priors that do not depend

*.
†.
‡.
§Research supported by an ISF grant.

on details of the experimental design or on knowing the specific functions of the parameters that may turn out to be of interest. In this paradigm, it would not, for example, be reasonable for a statistician to use one prior for estimating $h_1(\vartheta)$ and another to estimate $h_2(\vartheta)$, unless $h_1 \equiv h_2$.

We must necessarily approach matters from a "robustness" point of view and consider what happens if the parameter has probability 0 under the assumed prior, as would happen with all reasonable "atom-less" priors on continuous parameter spaces. That is, we study Bayesian procedures from a frequentist point of view in the tradition of Bernstein, von Mises and Le Cam, and more recently Cox (1993), Diaconis and Freedman (1998), Freedman (1993), Freedman (1999), and also of Bayarri and Berger (2004).

The extent to which the subjective aspect of data analysis is central to the modern Bayesian point of view is debateable. See the dialog between Goldstein (2006) and Berger (2006a) and the discussion of these two papers. However, central to any Bayesian approach is the posterior distribution and the choice of prior. Even those who try to reconcile Bayesian and frequentist approaches, cf. Bayarri and Berger (2004), tend to give, in the case of conflict, a stronger preference to inferences based on the posterior, rather than frequentist properties, cf. Berger (2006b).

Most early discussions of Bayesian analysis presented simple examples, e.g., $X \sim N(\vartheta, 1)$. In this case, a statistician might have clear *a priori* ideas about ϑ, and might well understand the implications of using one prior in place of another. Regardless, the data will eventually overwhelm the prior, and typically frequentist and Bayesian inference will coincide. The classical Bernstein-von Mises Theorem encapsulate this observation, see Le Cam and Yang (1990) or Lehmann and Casella (1998). Currently, Bayesian procedures are being applied to complex, high-dimensional models, e.g., those used in medical imaging. With a very high-dimensional parameter space (where laws of large numbers appear, "uniform" distributions are concentrated on shells, etc), it is very difficult to understand the implications of using a particular prior (in place of another). It is very difficult if not impossible to express subjective information about the model in a robust prior, and it is difficult to express this knowledge in a way that would support the data analysis and not dominate it. This is the situation we want to address in the current paper, and to some extent has already been considered by Cox (1993), Freedman (1993), and others.

Admittedly, there is a body of theory in the area, cf. Ghosal, Ghosh and van der Vaart (2000), Kleijn and van der Vaart (2006), and Bickel and Kleijn (2012), among others, giving specific conditions under which some finite dimensional intuition persists in higher dimensions. However, in this paper we emphasize how easily these conditions can be violated and the dramatic consequences of such violations.

It is argued "that selection of prior distributions will rarely follow the idealized scenario of being done without reference to the data or experimental structure. After all, models are often selected only after a careful examination of the data, so how could a prior on model parameters have been selected beforehand? ... Bayesian model selection can temper the 'desire' of the data to be

overfitted, by bringing in prior weights that can be assigned to models" (Berger (1985), pg. 284). We argue that, for Bayesian modeling to have good frequentist properties, one has to consider not only aspects of the experimental structure, but also the specifics of the parameter of interest, and the fine details of the design. In contrast to protection against overfitting, priors may cause underfitting (i.e., believing what the prior says, even when it is not supported by the data). A prior that fits the task, is not a prior. We are reminded of Groucho Marx's quote "Those are my principles, and if you don't like them ... well, I have others." In short, we argue that the only way to judge whether a prior is good, is to check the behavior of the resulting Bayesian estimator. We do not argue that there are no good Bayesian estimators. But we do argue that the arguments that justify their use cannot be Bayesian.

We use several examples to illustrate a number of issues. In Section 2, we replicate the results of Robins and Ritov (1997) in a missing data problem for which a simple (non-Bayesian) estimator is efficient, while the application of any prior which adheres to the strict paradigm discussed above forces us to implicitly estimate a infinite-dimensional parameter, and leads to an estimator of the parameter of interest with a slow rate of convergence. This argument is aimed primarily at estimators which are efficient in the frequentist sense, and failures of the strict likelihood principle; but if the Bernstein-von Mises phenomenon holds, the corresponding Bayesian estimators are necessarily efficient. In Section 3 we consider the partial linear model of Engle, Granger, Rice and Weiss (1986). In this case, if the nonparametric part of the model is smooth enough, the Bernstein-von Mises phenomenon holds and Bayesian estimators are efficient, under some conditions on the prior, but frequentist estimation can get away with significantly less smoothness in the nonparametric regression term.

Sections 4 and 5 deal with parameters in the Gaussian white noise model, $X_i = \beta_i + \varepsilon_i$, $i = 1, 2, \ldots, \varepsilon_1, \varepsilon_2, \ldots$ i.i.d. $N(0, 1/n)$. In Section 4 we show that for any Bayes prior concentrating on sequences of means which decay slowly enough, there exist linear functionals of the parameters whose posterior distribution does not converge to the truth at the $n^{-1/2}$ rate, whereas simple frequentist estimators always do. This phenomenon reflects among other things that it is quite possible to have parallel mutually independent sequences which have similar temporal correlation and exhibit strong cross-correlation. This somewhat surprising situation is discussed in Appendix A.

This result is coupled with a dramatic example of the failure of a plausible prior on images to produce a decent estimate of a linear parameter of the noisy image as opposed to a naive frequentist estimator.

In Section 5 we study a quadratic functional which behaves as the slope does in the partial linear model. We argue there that "natural" reference priors only work over a limited range.

In Section 6 we give an example in which Bayesian procedures which ignore the stopping time associated with the data generating process fail, while simple frequentist procedures continue to work. This demonstrates the danger of the classical principle that Bayesians need not pay attention to stopping times.

Throughout, we argue that the parameter values for which the Bayes pro-

cedures fail are not atypical. A version of Doob's consistency theorem does however hold. If there exist $\sqrt{n}$ consistent frequentist estimates, failure of $\sqrt{n}$ consistency of the posterior can only hold on sets of prior probability 0. This is demonstrated in Section 7, where we also review and discuss our findings.

2. Continuously Stratified Random Sampling

Robins and Ritov (1997) consider an infinite-dimensional model of continuously stratified random sampling in which one has n i.i.d. observations $W_i = (X_i, R_i, Z_i)$ with $X_i \in [0,1]^d$, $Z_i = R_i Y_i$, and $R_i, Y_i \in \{0,1\}$ and are conditionally independent given X_i, with $g(X) = \mathrm{E}(R|X)$ known and $h(X) = \mathrm{E}(Y|X)$ unknown. The parameter of interest is $\vartheta = \mathrm{E}(Y)$. For discussion of this model see also Wasserman (1998) and Harmeling and Toussaint (2007).

It is relatively easy to construct a reasonable estimate of ϑ. Indeed, the classical Horvitz-Thompson estimator, cf. Cochran (1977)

$$\hat{\vartheta} = n^{-1} \sum_{i=1}^{n} Z_i / g(X_i)$$

solves the problem nicely. Because,

$$\begin{aligned} \mathrm{E}\{RY/g(X)\} &= \mathrm{E}\left\{\mathrm{E}(R|X)\mathrm{E}(Y|X)/g(X)\right\} \\ &= \mathrm{E}\,\mathrm{E}(Y|X) = \vartheta, \end{aligned}$$

the estimator is consistent without any further assumptions. If we assume that g is bounded from below, the estimator is $\sqrt{n}$-consistent and asymptotically normal

Consider now a Bayesian analysis of the problem. To simplify the discussion, assume that the X_i are sampled from an absolutely continuous distribution F on $[0,1]^d$ with known density $f(x)$. As f and g are both known, the only remaining parameter is h, where $h(X) = \mathrm{E}(Y|X)$. Let π be a prior for h with respect to some measure μ. The joint density of h and the observations is given by

$$\begin{aligned} p(h,\mathbf{W}) = \pi(h) &\prod_{i\,:\,R_i=1} h(X_i)^{Y_i} \left(1 - h(X_i)\right)^{1-Y_i} \\ &\times \prod_{i=1}^{n} g(X_i)^{R_i} \left(1 - g(X_i)\right)^{1-R_i} \end{aligned}$$

as $Z_i = Y_i$ when $R_i = 1$. But this means that the posterior for h satisfies

$$\pi(h|\mathbf{W}) \propto \pi(h) \prod_{i\,:\,R_i=1} h(X_i)_i^{Y} \left(1 - h(X_i)\right)^{1-Y_i}. \tag{1}$$

Of course, this is only a function of those observations for which $R_i = 1$, for which the Y_i are directly observed; that is, the observations for which $R_i = 0$ are deemed uninformative. The difficulty with this restricted point of view is

quite simply that the Bayesian can only make use of the information contained in (1). However, (1) is independent of g. Hence, any procedure which depends on g, for example, the Horvitz-Thompson estimator, cannot be used in this analysis. The Bayesian is restricted to estimates of ϑ determined by estimates of h and, when d is large, estimating h can be very difficult. Indeed, if we assume that h is Hölder continuous with constants $M < \infty$ and $\alpha > 0$ (i.e., $\sup_{t>0} \sup_{0<u<1-t} t^{-\alpha}|h(u+t) - h(u)| < M$), we need $\alpha > d/2$ for our estimate of ϑ to be $\sqrt{n}$-consistent. If d is large, this is a very restrictive assumption.

If much is known about h, for example, there is a finite-dimensional parametric model for h, then the Bayesian paradigm runs into no particular difficulty. And, as above, similar claims can be made for less restrictive specifications. If, however, the problem is nonparametric and we wish to impose only minimal assumptions on h, the only available estimator is the Horvitz-Thompson estimator (or an estimator which is asymptotically very close to it), and such estimators are not available to the Bayesian nonparametric statistician.

Consider now the case in which neither g nor h is known, and these parameters are assumed *a priori* to be independent, with joint density $\pi(h)\rho(g)$ relative to some dominating measure. In this case, the posterior is the product of a term depending on g and a term depending on h, and information about g cannot be used by the Bayesian nonparametric statistician to construct estimates of ϑ. Unfortunately, it can be very difficult to construct reasonable estimates of g when X is high-dimensional.

If, under the prior, we assume g is sufficiently smooth, then the posterior can be used to obtain consistent estimates of g which in turn yield $\sqrt{n}$-consistent estimates of ϑ. But the rate at which g can be estimated is only $n^{-\alpha/(2\alpha+d)}$, where α is a measure of smoothness and d is the dimension. On the other hand, the frequentist Horvitz-Thompson estimator is efficient over all smoothness classes for g.

Regardless, if g is unknown, h cannot be estimated in general! This is true even in the one-dimensional case. Suppose X is distributed uniformly on the unit interval and g is given by

$$g(x) = \frac{1}{2} + \frac{1}{4} \sum_{i=0}^{m-1} s_i \, \psi\left(mx - i\right),$$

where $m = m_n = n^3$; the sequence $s_1, \ldots, s_m \in \{-1, 1\}$ is assumed to be exchangeable with $\sum s_i = 0$, and $\psi(x) = \mathbf{1}(0 \le x < 1/2) - \mathbf{1}(1/2 \le x \le 1)$. Furthermore, assume that $h(x) \equiv 17/64$ or $h(x) \equiv g(x)$. With probability converging to 1, there will be no interval of length $1/m$ with more than one X_i. However, given that there is one $X_i \in (j/m, (j+1)/m)$, then the distribution of (R_i, Z_i) is the same whether $h(x) \equiv 17/64$ or $h(x) \equiv g(x)$, and hence ϑ is not identifiable, and can be either $17/64$ or $1/2$.

Of course, in general, it is possible to trade smoothness in g for a lack of smoothness in h and vice versa, to construct estimates of ϑ but smoothness assumptions in high dimensions tend to be restrictive.

Note that this argument shows that any estimator which is based only on the likelihood function, ignoring auxiliary information which is not part of it or the parameter space, fails in this setup. In particular, this includes the maximum likelihood estimator.

However, we can construct a $\sqrt{n}$-consistent Bayesian estimator if it is based on the *a priori* unknown g. An example is given in Appendix B.

This argument mixes Bayesian and non-Bayesian techniques. Our goal is to make the argument precise and to study its impact on understanding the meaning of Bayesian inference in complex, high-dimensional models.

3. The Partial Linear Model

In this section we consider the partial linear model, also known as the partial spline model, and originally discussed in Engle et al. (1986); see also Schick (1986). In this case, we have observations $W_i = (U_i, X_i, Y_i)$ such that

$$Y_i = \vartheta X_i + g(U_i) + \varepsilon_i$$

where (X_i, U_i) are i.i.d. samples from the joint density $p(x, u)$, relative to Lebesgue measure on the unit square, $[0, 1]^2$; g is an element of some class of functions, $\mathcal{G}$; and the ε_i are i.i.d. $N(0, 1)$. The parameter of interest is ϑ and g is a (possibly very non-smooth) nuisance parameter. Let $h(U) = \mathrm{E}\left(X \,\middle|\, U\right)$. For simplicity, assume that U is known to be uniformly distributed on the unit interval.

3.1. A Frequentist Analysis

The loglikelihood function is

$$\ell(\vartheta, g, p) = -\frac{(y - \vartheta x - g(u))^2}{2} - \log p(x, u).$$

It is straight forward to argue that the score function for ϑ (the directional derivative of the log-likelihood in the least favorable direction for estimating ϑ) is given by (cf. Schick (1986); Bickel, Klaassen, Ritov and Wellner (1998))

$$\tilde{\ell}_\vartheta\left(\vartheta, g\right) = \left(x - h(u)\right)\left(y - \vartheta x - g(u)\right) = \left(x - h(u)\right)\varepsilon,$$

and the semiparametric information bound for the estimation of ϑ is

$$I = \mathrm{E}\,\mathrm{Var}(X|U).$$

We assume that $I > 0$. In particular, this implies that X is not a function of U.

Under some regularity conditions an efficient estimator can be constructed along the following lines. Find initial estimators $\tilde{h}$ and $\tilde{g}$ of h and g respectively, and estimate ϑ by computing

$$\hat{\vartheta} = \frac{\sum\left(X_i - \tilde{h}(U_i)\right)\left(Y_i - \tilde{g}(U_i)\right)}{\sum\left(X_i - \tilde{h}(U_i)\right)^2}.$$

The idea is that ϑ is the covariance between X and Y conditional on any given values of U and this estimator is based on the assumptions that the conditional expectation of X and Y given U are smooth enough, and have a fair estimators $\tilde{h}$ and $\tilde{g}$ respectively.

We could, for example, assume that the functions g and h satisfy Hölder conditions of order α and and γ, respectively. That is, there is $C < \infty$ such that $|g(v) - g(u)| \leq C|v-u|^{\alpha}$ and $|h(v) - h(u)| \leq C|v-u|^{\gamma}$ for all v, u in the support of U. We could also assume that $\mathrm{Var}(X|U)$ has a version which is continuous in u. In this case, so long as $\alpha+\gamma > 1/2$, and $I > 0$, we can construct a $\sqrt{n}$-consistent and semiparametrically efficient estimate of ϑ.

An estimator that tries to push the smoothness assumptions to the absolutely weakest necessary is the following:

Let $W_{(i)} = (U_{(i)}, X_{(i)}, Y_{(i)})$, $i = 1, 2, \ldots, n$ be the sample ordered such that $U_{(1)} \leq U_{(2)} \leq \ldots$. Write $X = h(U) + Z$. Note that $\sum \left(U_{(i+1)} - U_{(i)}\right)^2 = \mathcal{O}_P\left(n^{-1}\right)$ under the assumption that the U_i are uniformly distributed on $[0,1]$, while $n^{-1}\sum \left(X_{(i+1)} - X_{(i)}\right)^2 \xrightarrow{\mathrm{p}} c > 0$ under the assumption that $I > 0$. Take

$$\begin{aligned}\tilde{\vartheta} &= \frac{\sum \left(X_{(i+1)} - X_{(i)}\right)\left(Y_{(i+1)} - Y_{(i)}\right)}{\sum \left(X_{(i+1)} - X_{(i)}\right)^2} \\ &= \vartheta + \frac{\sum \left(X_{(i+1)} - X_{(i)}\right)\left(\varepsilon_{(i+1)} - \varepsilon_{(i)}\right)}{\sum \left(X_{(i+1)} - X_{(i)}\right)^2} + R,\end{aligned}$$

where

$$\begin{aligned}R &= \frac{\sum \left(X_{(i+1)} - X_{(i)}\right)\left(g(U_{(i+1)}) - g(U_{(i)})\right)}{\sum \left(X_{(i+1)} - X_{(i)}\right)^2} \\ &= \frac{\sum \left(Z_{(i+1)} - Z_{(i)}\right)\left(g(U_{(i+1)}) - g(U_{(i)})\right)}{\sum \left(X_{(i+1)} - X_{(i)}\right)^2} \\ &\quad + \frac{\sum \left(h(U_{(i+1)}) - h(U_{(i)})\right)\left(g(U_{(i+1)}) - g(U_{(i)})\right)}{\sum \left(X_{(i+1)} - X_{(i)}\right)^2} \\ &= o_P\left(n^{-1/2}\right) + O_P\left(n^{-(\alpha+\gamma)}\right).\end{aligned}$$

because the Z_i are uncorrelated with the U_i. On the other hand,

$$\sqrt{n}\frac{\sum \left(X_{(i+1)} - X_{(i)}\right)\left(\varepsilon_{(i+1)} - \varepsilon_{(i)}\right)}{\sum \left(X_{(i+1)} - X_{(i)}\right)^2} \xrightarrow{\mathrm{D}} N\left(0, \frac{3}{2}I^{-1}\right).$$

We conclude that ϑ can be estimated in a $\sqrt{n}$ rate if h is Lipschitz of order γ, g Lipschitz of order α, and $\alpha + \gamma > 1/2$. This estimator is not efficient, but it does show what the minimal local smoothness conditions are. We want to remark that not all pairs of observation are needed, a subset of size of order n may suffice.

For the sake of completeness, we construct an efficient estimator. Let $c_n \to 0$ slowly, and choose three random sub-samples of size $c_n n$. Based on the first sub-sample construct an initial estimate of ϑ; estimate g using the second; and h using the third. Denote these estimates $\tilde{\vartheta}$, $\tilde{g}$, and $\tilde{h}$, respectively. The non-parametric components g and h are identified as $E(Y - \vartheta X|U)$ and $E(X|U)$. Both can be estimated using kernels (with $\tilde{\vartheta}$ plugged in for ϑ) with bandwidth $\sigma_n \to 0$ slowly enougth that $c_n \sigma_n n \to \infty$. Then $E(\tilde{g} - g)^2 = \mathcal{O}(n^{-2\alpha+2\nu})$, $E(\tilde{h} - h)^2 = \mathcal{O}(n^{-2\gamma+2\nu})$, with ν arbitrary small. Let S be the remainder of the sample. Calculate

$$\begin{aligned}
\hat{\vartheta} &= \frac{\sum_{i\in S}\big(X_i - \tilde{h}(U_i)\big)\big(Y_i - \tilde{g}(U_i)\big)}{\sum_{i\in S}\big(X_i - \tilde{h}(U_i)\big)^2} \\
&= \frac{\sum_{i\in S}\big(X_i - \tilde{h}(U_i)\big)\big(\vartheta X_i + g(U_i) + \varepsilon_i - \tilde{g}(U_i)\big)}{\sum_{i\in S}\big(X_i - \tilde{h}(U_i)\big)^2} \\
&= \vartheta + \frac{\sum_{i\in S}\big(X_i - h(U_i)\big)\varepsilon_i}{\sum_{i\in S}\big(X_i - \tilde{h}(U_i)\big)^2} + \frac{\sum_{i\in S}\big(h(U_i) - \tilde{h}(U_i)\big)\varepsilon_i}{\sum_{i\in S}\big(X_i - \tilde{h}(U_i)\big)^2} \\
&\quad + \frac{\sum_{i\in S} Z_i\big(g(U_i) - \tilde{g}(U_i)\big)}{\sum_{i\in S}\big(X_i - \tilde{h}(U_i)\big)^2} + \frac{\sum_{i\in S}\big(h(U_i) - \tilde{h}(U_i)\big)\big(g(U_i) - \tilde{g}(U_i)\big)}{\sum_{i\in S}\big(X_i - \tilde{h}(U_i)\big)^2} \\
&= \hat{\vartheta}^* + \circ_P(n^{-1/2}) + \mathcal{O}_P\left(n^{-1}\sum_{i\in S}\big(h(U_i) - \tilde{h}(U_i)\big)\big(g(U_i) - \tilde{g}(U_i)\big)\right).
\end{aligned}$$

since the initial estimators are independent and independent of S. We conclude that $\hat{\vartheta} = \vartheta + \sum_i \big(X_i - h(U_i)\big)\varepsilon_i + \circ_p(n^{-1/2})$, as required.

3.2. Minimal Smoothness

Consider the sub-model where $(\vartheta, g, h) \in \mathbb{R} \times \mathcal{A}_m \times \mathcal{A}_m$:

$$\begin{aligned}
\mathcal{A}_m = \{h : h(u) = \sum_{i=0}^{m} c_i (v_{i+1} - v_i)^\alpha \psi\Big(\frac{u - v_i}{v_{i+1} - v_i}\Big), \\
u \in (0,1), \quad 0 = v_0 < v_1 < \cdots < v_{m+1} = 1, \quad c_i \in \mathbb{R}, \max|c_i| \le M\}
\end{aligned}$$

where

$$\psi(u) = t^\alpha \mathbf{1}_{[0,1/2)}(u) - (1-t)^\alpha \mathbf{1}_{[1/2,]}(u).$$

Clearly, if $(g, h) \in \mathcal{A}_m \times \mathcal{A}_m$, they are Hölder of order α. Suppose $n \max\{u_{i+1} - u_i\} \to 0$, then the constants $c_1, \ldots, c_m$ that define g and h cannot be estimated consistently, and hence if the v's appear behave like a size m random sample from a uniform distribution, $\|\hat{h} - h\|^2 = \mathcal{O}_p(m^{-\alpha})$.

Consider now a semi-Bayesian version of the problem, where $m = n^{1+\nu}$, $\nu > 0$, $v_1, \ldots, v_m$ are the order statistics of a sample of i.i.d. $U(0,1)$, $c_1, \ldots, c_m$

are independent, the c_i in the definition of $\mathcal{A}_m$ are $N(0,\tau^2)$ for h and $N(0,\eta^2)$ for g, with correlation ρ between the two, and $Z = X - h(U)$ is a $N(0,\nu^2)$ random variable. Finally, $\varepsilon_i \sim N(0,\sigma^2)$. Then we observe pairs

$$\begin{aligned} X_i &= h_i + Z_i \\ Y_i &= \vartheta(h_i + Z_i) + g_i + \varepsilon_i. \end{aligned}$$

Hence they follow a bivariate normal distribution with covariance matrix:

$$\begin{bmatrix} \nu^2+\tau^2 & \vartheta(\nu^2+\tau^2)+\rho\tau\eta \\ \vartheta(\nu^2+\tau^2)+\rho\tau\eta & \sigma^2+\eta^2+\vartheta^2(\nu^2+\tau^2)+2\vartheta\rho\tau\eta \end{bmatrix}.$$

Any estimator of the estimator is equivalent to solving the empirical covariance matrix (which yields 3 equations) for the 6 parameters $(\vartheta, \nu^2, \sigma^2, \rho, \tau, \eta)$. If $\alpha < 1/4$ then $\tau^2, \eta_2 \gg n^{-1/2}$, hence the expression $\rho\tau\eta$ cannot be ignored, and ϑ cannot be solved to the $n^{-1/2}$ accuracy.

3.3. A Bayesian Analysis

We want to consider a Bayesian approach. Suppose that the Bayesian has an independent priors on $p(u,x)$, g and ϑ, $\pi = \pi_p \times \pi_g \times \pi_\vartheta$. For example, the first distribution may be a function of the environment, the prior on the non-parametric component of the regression function is a function of the physical process and the third component of the prior is about our understanding of the measurement engineering. The log-posterior is then

$$A - \frac{1}{2}\sum_{i=1}^{n}\Big(Y_i - \vartheta X_i - g(U_i)\Big)^2 + \pi_\vartheta(\vartheta) + \pi_g(g) + \sum_{i=1}^{n}\log p(u,x) + \pi_p(p).$$

That is. when the Bayesian comes to estimate ϑ, he does not see any information about h. The same estimator would be used whatever is known about the smoothness of h!

Suppose now that essentially it is only known that g is Hölder of order α, while the range of U is divided to some intervals, such that h is either Hölder of order γ_1 or of order γ_2 where

$$\alpha + \gamma_1 < \frac{1}{2} < \alpha + \gamma_2.$$

Then a $\sqrt{n}$ consistent estimator should only use the intervals where h is Hölder of order of γ_2. The rest should be discarded. If the number of observations in the "good" intervals is of the same order as n, then the estimator is still $\sqrt{n}$ consistent. For a frequentist, there is no difficulty in ignoring the nuisance intervals. ϑ is assumed to be the same all over. However, the Bayesian cannot ignore these intervals. In fact, his *a posteriori* distribution does not contain any information which intervals are good and which are bad.

There is no logical contradiction. The type of parameters combination that the Bayes estimator fails on, has negligible *a priori* probability. He assumes

that the *a priori* g and h are independent, and short intervals are essentially independent (to take care of the very rough g and h we deal with). Under these assumptions, the intervals with h Hölder of order γ_1 contribute on the average 0. But this average is by the prior, which was conveniently constructed by the Bayesian.

4. The white-noise-model and the plug-in property of Bayesian estimate

We consider the white noise model:

$$X_i = \beta_i + \varepsilon_i,$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots) \in \mathcal{B} \subset \ell_2$, and $\varepsilon_1, \varepsilon_2, \dots$ are i.i.d. $N(0, 1/n)$. This model is called the white noise model because its equivalence to the model $dX(t) = \mu(t) + n^{-1/2} dW(t)$, $t \in [0,1]$, $\mu \in L_2$, and W a standard Wiener process, by taking $X_1, X_2, \dots$ and $\beta_1, \beta_2, \dots$ to be the projection of $X(\cdot)$ and $\mu(\cdot)$ on some orthonormal basis of $L_2(0,1)$. We consider the estimation of $\boldsymbol{\beta}$ as an object of ℓ_2 with a squared norm loss function $\|\hat{\beta} - \boldsymbol{\beta}\|^2$, and estimation of a linear functional of $\boldsymbol{\beta}$, $h(\boldsymbol{\beta}) = \sum_{i=1}^{\infty} c_i \beta_i$, $h \in \mathcal{H} = \{h(\boldsymbol{\beta}) = \sum_{i=1}^{\infty} c_i \beta_i : \ (c_1, c_2, \dots) \in \ell_2\}$, again under the error squared loss function.

To be more specific we consider $\mathcal{B}_\alpha = \{\boldsymbol{\beta} : |\beta_i| < i^{-\alpha}\}$, $\alpha > 1/2$. From a standard frequentist point of view the estimation in this problem is simple enough. Simple estimators that achieve the optimal rate of convergence are given in the following proposition:

Proposition 4.1 *The estimator $\hat{h} = \sum h_i X_i$ is $\sqrt{n}$ consistent for any $h \in \mathcal{H}$. The estimator*

$$\hat{\beta}_i = \begin{cases} X_i & i^\alpha \le n^{1/2} \\ 0 & i^\alpha > n^{1/2} \end{cases}$$

achieves the minimax rate of convergence, $n^{-(2\alpha-1)/2\alpha}$.

The proof is in Appendix C.

A major characterization of the Bayes procedures is that they have necessarily the plug-in-property (PIP). Since

$$\mathrm{E}\, h(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{\infty} c_i \mathrm{E}\, \hat{\boldsymbol{\beta}},$$

we have $\widehat{h(\boldsymbol{\beta})} = h(\hat{\boldsymbol{\beta}})$, for any Bayes estimators of $h(\boldsymbol{\beta})$ and $\boldsymbol{\beta}$, respectively, both under quadratic loss function.

However, there is no efficient estimator with PIP in the white noise model as is shown in Bickel and Ritov (2003). Every estimator would fail either as a nonparametric estimator with an optimal rate, or as a plug-in-estimator (PIE) of at least one linear functional. The argument of Bickel and Ritov (2003), being

valid to any estimator is not strong enough for our purpose. However, we can strengthen it for Bayes procedures.

We need the following lemma, whose proof is given in Appendix C.

Lemma 4.2 *Suppose* $X \sim N(\vartheta, \sigma^2)$, $|\vartheta| \leq a \leq \sigma$. *Let* $\hat{\vartheta} = \hat{\vartheta}(X)$ *be the* a posteriori *mean when the prior is* π. *Let* b_ϑ *be its the bias under* ϑ. *Then* $|b_\vartheta| + |b_{-\vartheta}| > 2(1-(a/\sigma)^2)|\vartheta|$. *In particular, if* π *is symmetric around 0, then* $|b_\vartheta| > (1-(a/\sigma)^2)|\vartheta|$.

The proof is in Appendix C.

This lemma shows that any Bayes estimator is necessarily biased and puts a lower bound on this bias. We will use this lemma to argue that any Bayes estimator is going to fail for some simple functionals.

Theorem 4.3 *For any Bayesian estimator* $\hat{\boldsymbol{\beta}}$ *with respect to prior on* $\mathcal{B}_\alpha$, $\alpha > 1/2$, *there are* $h \in \mathcal{H}$ *and* $\boldsymbol{\beta} \in \mathcal{B}_\alpha$ *such that* $n\big(h(\hat{\boldsymbol{\beta}}) - h(\boldsymbol{\beta})\big)^2 \xrightarrow{\mathrm{p}} \infty$. *In fact,* $\mathrm{E}_{\boldsymbol{\beta}}\, h(\hat{\boldsymbol{\beta}}) - h(\boldsymbol{\beta}) = \mathcal{O}(n^{-2\alpha-1/4\alpha})$.

Proof. It follows from Lemma 4.2 that for any $i > 2n^{1/2\alpha}$ there is β_i such that if $b_i = \mathrm{E}\,\hat{\beta}_i - \beta_i$ then $|b_i| > 3i^{-\alpha}/4$. Define

$$c_i = \begin{cases} 0 & i \leq 2n^{1/2\alpha} \\ C_1 n^{(2\alpha-1)/4\alpha}\, i^{-\alpha} & i > 2n^{1/2\alpha} \;\&\; b_i > i^{-\alpha}/2 \\ -C_1 n^{(2\alpha-1)/4\alpha}\, i^{-\alpha} & i > 2n^{1/2\alpha} \;\&\; b_i < -i^{-\alpha}/2, \end{cases}$$

where C_1 ensures that $\sum_{i=1}^{\infty} c_i^2 = 1$ (note that C_1 is bounded away from 0 and ∞). Hence

$$\begin{aligned} \mathrm{E} \sum c_i(\hat{\beta}_i - \beta_i) &\geq \frac{3}{4} C_1 n^{(2\alpha-1)/4\alpha} \sum_{i>2n^{1/2\alpha}} e^{-2\alpha} \\ &\geq \frac{3}{4} C_1 n^{-(2\alpha-1)/4\alpha}. \end{aligned}$$

□

That is, any Bayesian estimator fails on some pairs β and h. These pairs are not strange animals. Actually they are pretty 'typical' members of $\mathcal{B}_\alpha$ and $\mathcal{H}$. What makes them special is only that the sequence $\beta_1, \beta_2, \ldots$ is not ergodic, and similarly $h_1, h_2, \ldots$ is not. Each of them have a non-trivial auto-correlation function, and the two auto-correlation functions are similar. The prior makes such pairs unlikely, and the biases of the estimator of each of the components are going to cancel each other by the prior. If the *a priori* distribution is presenting a real physical phenomena, this exact cancelation, due to the law of large numbers is reasonable, and the statistician should not worry about it. If the prior is a way to express ignorance, or beliefs—subjective beliefs—than one should worry about these small biases. Certainly so, if the only reason to assume that small terms are not going to accumulate is based on mathematical convenience of expressing rough ideas about the unknown parameters.

The parameter $\boldsymbol{\beta}$ and the functional may be similar because of phenomena such as the one presented in Appendix A. In a large space, the autocorrelation function may be complex with an unknown neighborhood structure, and in practice, completely hidden from the observer.

We consider a Bayesian model to be honestly nonparametric on $\mathcal{B}_\alpha$, if $\mathcal{L}(\beta_i \mid X_{-i})$ is symmetric around 0, and $P(\beta_i > \gamma i^{-\alpha} \mid X_{-i}) > \gamma$, for some $\gamma > 0$, where $X_{-i} = X_1, \ldots, X_{i-1}, X_{i+1}, \ldots$. That is, at least in some sense, the components of β_i are free parameters. We have:

Theorem 4.4 *Suppose the prior is honestly non-parametric on $\mathcal{B}_\alpha$ and $1/2 < \alpha < 3/4$, then the Bayesian estimator of $h(\boldsymbol{\beta}) = \sum_{i=1}^{\infty} c_i\beta_i$ is not consistent, if $|c_i| = i^{-\alpha}$, $\boldsymbol{\beta}$ and h are serially correlated with bounded away from 0 correlation, and $\sqrt{m}\sum_{i=m}^{\infty}\beta_i^2 \to \infty$ (which is the a.s. the case under the prior).*

Proof. Again, we consider the bias as in the second part of Lemma 4.2:

$$\sqrt{n}\mathrm{E}\sum_{i>n^{\nu+1/2\alpha}} c_i(\hat{\beta}_i - \beta_i) = \sum_{i>n^{\nu+1/2\alpha}} d_i c_i \beta_i, \qquad |d_i - 1| < n^{-\nu}.$$

□

4.1. An example

Here is a simple simulation. For the vector $\boldsymbol{\beta}$ we considered the image given in Figure 1(a): the figure is a gray scale image of a 367×300 matrix, whose vectorization is the vector $\boldsymbol{\beta} \in \mathbb{R}^{110,100}$. That is, if the gray level of the image represents I_{ij}, then $\beta = \vec{I} = (I_{1,1}, I_{2,1}, \ldots, I_{367,1}, I_{1,2}, \ldots, I_{367,300})$. To obtain X we added to each pixel an independent $N(0, 169)$ random variable. See Figure 1(b). We emphasis that we do not consider $\boldsymbol{\beta}$ and X as images, but as vectors with exchangeable components. The Bayes estimator was calculated with respect to prior which considers the components as i.i.d. $N(\mu, \tau^2)$, where $\mu = \sum w_i\beta_i / \sum w_i$, $w_1, w_2, \ldots$ are i.i.d. $U(0, 1)$ random variables and $\tau^2 = 315.786$ is the true empirical variance of $\beta_1, \ldots, \beta_{110100}$. The resulted SNR is low (-2.72db). The nonparametric Bayes estimator of the $\boldsymbol{\beta}$ is given in Figure 1(c). It is closer to the true image than the noisy observations, as expected, since the prior is a honest exchangeable description of the data.

The purpose of the estimation was not the nonparametric estimate per se, but in the spirt of this section, an estimation of a functional. The functional h is given in Figure 1(d). Again, the image is a gray scale representation of $c_1, \ldots, c_{110100}$. The two images were selected from the small collections of images supplied by the standard distribution of Matlab, and this pair was selected because their sizes fitted. Thus we have two processes on the unit square. One represents $\{\beta_i\}$, while each pixel in the jet image, represents the value of h_i. In both cases, the image is an image of an object, and therefore has a center and margins. There is a strong correlation between the point of the picture, above being continuous. The vector of parameters and the stopping times are correlated, being referred

to images that follow the same rules of good image, rules hat are not necessarily known to the data analyst.

Applying h to the noisy X yields an estimate with RMSE (root mean squared error) of 1.04. Applying h to the much cleaner Bayes estimator gives RMSE of 19.01. The main difference between these two estimators was in the bias (0.01 versus 19.00). The bias and RMSE calculation were based on 500 Monte Carlo simulations.

The Bayes estimator does not fail because the object of interest, $\boldsymbol{\beta}$, and the functional are not independent. They are independent. There is no reason to assume that the bone structure of the image representing the functional has anything to do with the jet imaged in the object. It did not fail because no image analysis tools were used. Smoothness of the picture is far from being relevant to the failure. They failed because the prior failed to recognize that the images are not permutation invariant or ergodic, and hence two images may be correlated, positively or negatively by chance, but correlated. See Appendix A. In fact, this is typically the case with two good pictures. A good pictures has a structure. It has a center and it has margins. it is not a mixing process. Now, with pictures this is easy to understand in retrospect. Not that easy to understand *a priori*. But pictures are two dimensional and at least can be viewed. With complicated graphs, which human beings cannot viewed and understand, but may be non-mixing and with clear (not well understood) structure, the same situation could happen, bias would be introduced into the Bayes procedure, but the Bayesian may fail to understand, and the prior that expresses his subjective belief on the subject would fail to protect him against bias.

5. Estimating the signal squared, and the importance of being unbiased

We continue with the analysis of the white noise model of Section 4, but we consider a different Euclidean parameter of interest: $\vartheta = \sum_{i=1}^{\infty} \beta_i^2$.

A natural estimator of β_i is given by Proposition 4.1, and one may consider as an estimator of the parameter $\tilde{\vartheta} = \sum \tilde{\beta}_i^2 = \sum_{i<n^{1/2\alpha}} X_i^2$. This works fine when $\alpha > 1$. It achieves both the minimax rate for estimating $\boldsymbol{\beta}$, and $\tilde{\vartheta}$ is an efficient estimator of the Euclidean parameter. But $\tilde{\beta}^2$ has a bias of n^{-1} as an estimator of β_i^2, which accumulate to $n^{-1+1/2\alpha} \gg n^{-1/2}$ when $\alpha < 1$. The simple traditional correction is to unbias the estimator, cf. Bickel and Ritov (1989):

Proposition 5.1 *Suppose $\alpha \in (3/4, 1)$, then an efficient estimator of ϑ is given by*

$$\hat{\vartheta} = \sum_{i \le m} (X_i^2 - \frac{1}{n}),$$

for $n^{1/(4\alpha-2)} < m \ll n$.

Proof. Clearly the bias of the estimator is bounded by $\sum_{i>m} i^{-2\alpha} < m^{-(2\alpha-1)} = \circ_p(n^{-1/2})$, and its variance is bounded by $\sum_{i\le m}(4\beta_i^2/n + 2/n^2) = 4\vartheta + \circ_p(n^{-1})$. □

However, this is a standard frequentist approach. There is a problem, and a solution is justified because it works, and not because it fits a paradigm. The solution works because we see that bias accumulation is the issue and we can deal with it.

5.1. The Bayesian analysis

Consider the above situation with $\alpha \in (3/4, 1)$. Then the estimator suggested in Proposition 5.1 sums at least $n^{1/(4\alpha-2)}$ terms. Note that most of the terms, all beyond the first $n^{1/2\alpha}$, are deeply under the noise level! This creates a problem for Bayesian analysis.

For any prior on β_i, $i > n^{1/2\alpha+\nu}$ with ν as small as needed and $m = n^{1/(4\alpha-2)}$ as in Proposition 5.1:

$$\mathrm{E}_{\pi_i}(\beta_i^2 \mid X_1, \ldots, X_m) = \frac{\int_{-i^{-\alpha}}^{i^{-\alpha}} t^2 \varphi(n(X_i - t)) d\pi_i(t)}{\int_{-i^{-\alpha}}^{i^{-\alpha}} \varphi(n(X_i - t)) d\pi_i(t)} \in (A_i^{-1}\mathrm{E}_{\pi_i}\beta_i^2, A_i \mathrm{E}_{\pi_i}\beta_i^2)$$

where

$$\max_{n^{1/2+\nu}<i\le n^{1/4\alpha-2}} \log A_i \le \max_{\substack{n^{1/2+\nu}<i\le n^{1/(4\alpha-2)} \\ |t_i|<i^{-\alpha}}} \frac{n}{2}|(X_i - t_1)^2 - (X_i - t_2)^2| \xrightarrow{\mathrm{p}} 0,$$

since $\max_{i\le n}|X_i| n^{-1/2-\nu} \xrightarrow{\mathrm{p}} 0$. But this means that all the tail of β_i^2 for $i > n^{1/2\alpha}$ is replaced essentially by its *a priori* mean. Since this tail may carry signal of order $n^{-(2\alpha-1)/2\alpha} \gg n^{-1/2}$, the Bayes estimator is not consistent.

Where is the difference between the Bayes estimator and the frequentist estimator of Proposition 5.1? Both try to be unbiased. For the frequentist this mean that whatever is the value of the parameter, the estimator has expectation which is very close to the estimated parameter. The Bayesian however is unbiased with respect to his prior. Thus it is easy to him to replace whatever is difficult to estimate by its expectation according to the prior. This makes his estimator inconsistent in the frequentist sense, and inconsistent in any regular sense if the estimator does not describe exactly the generating mechanism of the data.

6. Data dependent sample size

The stopping rule principle says roughly that Bayesian inference should not depend on any stopping rule used to obtain the data to be analyzed, as long as it was done using stopping times. Formally, Berger and Wolpert (1988) wrote:

> "*Stopping Rule Principle (SRP)*: In a sequential experiment E^τ, with observed final data $\boldsymbol{x}^n$, $Ev(E^\tau, \boldsymbol{x}^n)$ should not depend on theThe proof is in Appendix C. stopping rule τ."

We challenge how this principle works with high dimensional data. We consider another version of the white noise model. We consider a finite version of it, with $n^{-2\alpha} < \beta_i < 3n^{-2\alpha}$, $i = 1, \ldots, k = \lfloor n^{2\alpha} \rfloor$, and $1/6 < \alpha < 1/4$. The ith component is a Brownian motion with drift β_i, $X_i(\cdot)$ is observed until time T_i. Let $\bar{X}_i(t) = X_i(t)/t$, the sufficient statistic for β_i given $\{Y_i(s) : s < t\}$. Of course, $\bar{X}_i$ is also the MLE. Finally, let $\boldsymbol{\pi}_i$ be the prior distribution of β_i given X_{-i}, the set of all observed components other than X_i. Let $\boldsymbol{f}_i(\cdot)$ be the distribution of $\bar{X}_i(T_i)$ given X_{-i} (i.e., $\boldsymbol{f}_i = \boldsymbol{\pi}_i * N(0, 1/T_i)$). We assume that the prior is nonparametric in the sense that $\boldsymbol{\pi}_i$ is bounded away from 0 on the permitted support, thus the rest of the data does not reveals too much on β_i.

It is well known that the posterior mean of β_i satisfies

$$\mathrm{E}(\beta_i \mid data) = \bar{X}_i(T_i) + \frac{1}{T_i} \frac{\boldsymbol{f}_i'(\bar{X}_i(T_i))}{\boldsymbol{f}_i(\bar{X}_i(T_i))}.$$

If $T_i = \mathcal{O}_p(n)$, then $\boldsymbol{f}_i \approx \boldsymbol{\pi}_i$. Further, $\bar{X}_i(T_i) \approx \beta_i$. Suppose T_i is correlated with $\boldsymbol{f}_i'/\boldsymbol{f}_i(\beta_i)$, then the MLE of $\sum_{i=1}^k \beta_i$, $\sum_{i=1}^k \bar{X}_i(T_i)$, has a random error of order $n^\alpha n^{-1/2}$, while the Bayes estimator has a bias which is $\mathcal{O}_p(n^{2\alpha} n^{2\alpha})/n)$ (there are $n^{2\alpha}$ terms, each one of them of size $n^{2\alpha}$ due to $T_i^{-1} \boldsymbol{f}_i'/\boldsymbol{f}_i$, and a factor of $1/n$ due to T). In the range of α we consider, the Bayes bias dominates the random error!

Consider now the stopping time:

$$T_i = \inf\{t : X_i(t) = n\beta_0 A_i + Z_i \sqrt{t}\},$$

where Z_i is a $N(0, 1)$ variable, A_i is an independent variable, whose values are under the control of an adversary, who is ready to tell their values to the statistician, but if the latter is a Bayesian, he simply ignore the former. The adversary is only restricted to have $E(T_i|A_i) = \Omega_p(n)$, which is the case if $A_i = \mathcal{O}_p(1)$.

All agree that $(T_i, X_i(T_i))$ are sufficient. In fact, the situation is more extreme. The distribution of $X_i(T_i)|T_i$ is independent of β_i, and hence either T_i or $X_i(T_i)$ is sufficient by itself!

This is a trivial statement for the frequentist who knows A_i. The Bayesian, cannot distinguish between this stopping time, and the situation in which T_i is endogenous, and the distribution of β_i given $T_i = t$ is the distribution of $n\beta_0 A_i/t$ (e.g., a point mass). Alternatively, his estimator is biased whenever, A_i is empirically correlated with β_i.

6.1. Example

We consider again the same vector $\boldsymbol{\beta}$ represented in Figure 1(a). But this time the spine image of 1(d) is giving the sample size per component. Noise was

added to obtain Figure 2(a). The SNR now, as can seen, is much higher than before (+2.72db). As a result the Bayes estimator given in Figure 2(b) is much smoother.

The prior was again with independent normal components with mean equals to the mean value of $\boldsymbol{\beta}$, and variance to its true empirical variance. Each pixel in the image was observed until a stopping time which was proportional to the gray level of the corresponding pixel in the spine image, Figure 1(d). Thus we have two processes on the unit square. One represents $\{\beta_i\}$, where each pixel in the jet image, represents the value of β_i. The process is of the stopping time, and is given by the spine image. In both cases, the image is an image of an object, and therefore has a center and margins. There is a strong correlation between the point of the picture, above being continuous. The vector of parameters and the stopping times are correlated, being referred to images that follow the same rules of good image, rules hat are not necessarily known to the data analyst. In 500 Monte Carlo simulations the RMSE of the mean Bayes estimate was 0.05 compared to the mean of the MLE RMSE which waas 0.009. The difference was almost all because of the bias. If we replace the stopping time with a fixed time, the average of the above, then the Bayes estimator is slightly better (RMSE of 0.0071 versus 0.0072). Thus, the example justifies the claim that the Bayes estimator failed when the stopping rule and the parameters values happened to be serially correlated.

7. A positive result and a summary

We start with a version of the Doob's consistency result, which shows that the existence of a uniform $\sqrt{n}$ consistent estimator ensures that the posterior distribution is $\sqrt{n}$ consistent with prior probability 1.

To simplify notation we consider in this section the Markov chain $\eta_0 \to X_n \to \eta_n$, where $\eta_0, \eta_n \in \mathcal{H}$, $\eta_0 \sim \pi$, $X_n \sim P_{\eta_0}$, and given X_n, η_0 and η_n are i.i.d., i.e., given X_n, η_n is distributed according to the *a posteriori* distribution π_{X_n}. Let P be the joint distribution of the chain. In the following d_n is a semi-metric on the parameter space, normalized to the sample size. Typically, in the nonparametric situation considered in this paper, $d_n(\eta, \eta') = \sqrt{n}|\vartheta(\eta) - \vartheta(\eta')|$ for some real functional ϑ of the parameter.

We consider an estimator $\tilde{\eta}_n$ to be d_n *consistent uniformly on* $\mathcal{H}$, if for all $\varepsilon > 0$ there is $M < \infty$ such that for all $\eta \in \mathcal{H}$ and n large enough, $P_\eta(d_n(\tilde{\eta}_n, \eta) \geq M) \leq \varepsilon$. The posterior is d_n *consistent uniformly on* $\mathcal{H}$ if for all $\varepsilon > 0$ and $\delta > 0$, there is $M < \infty$ such that for all $\eta_0 \in \mathcal{H}$ and n large enough, $P_{\eta_0}\big(\pi_{X_n}(d_n(\eta_n, \eta_0)) \geq M) \geq \varepsilon\big) \leq \delta$.

Theorem 7.1 *Suppose there exist a d_n consistent uniformly on $\mathcal{H}$ estimator. Then there is a $\mathcal{H}' \subseteq \mathcal{H}$ such that $\pi(\mathcal{H}') = 1$ and the posterior is d_n consistent on $\mathcal{H}'$.*

The proof is given in Appendix C.

Thus, the existence of a uniformly good frequentist estimator ensures that the Bayes posterior is concentrated in the right rate under all parameter values that seemed relevant under the prior. This claim does not contradict our findings. In the CODA and PLM examples, the difference between the Bayes estimator and the frequentist one, is that the former ignores the information that restricts the model to a subset of prior probability 0. In the quadratic function of the white noise model, the demand of the prior of being "honestly non-parametric", limited $\beta_1, \beta_2, \ldots$ to regular sequences obeying LLN's, and hence any non-ergodic sequence are in a set with probability 0. Finally, in the linear functional example, each prior fails for each linear functional on a set of parameters with probability 0, but if the linear functional and the parameter are chosen together as we argue it may happen, the theorem has no consequences.

In this paper we presented a few examples in which a nonparametric prior fails to estimate simple parametric functions at rate $n^{-1/2}$ even though frequentist efficient procedures exist. In this examples the assumed smoothness was minimal, but we do not believe that this is essential. With minimal smoothness it was easy to prove that the error explode. With smoother objects it would be more difficult to prove and estimators would be just not optimal.

Bayes procedure are always unbiased with the respect to the prior they are based upon. The Bayes estimator tends to replace elements buried inside the noise with their *a priori* mean. This would be a reasonable strategy if the prior represents a physical reality. If the prior represents subjective belief, not to say, a subjective belief based upon the need to have a prior that can be handled easily for highly plausible values of the parameters.

What were the phenomena that were exemplified in our models?

1. Spurious correlation. A possible empirical cross-correlation between two independent processes. The Bayes estimator ignores it. This happened in the CODA example of Section 2, the partial linear model of Section 3, the linear estimator of Section 4, and the stopping time story in Section 6. Since this correlation has expectation 0, the Bayes estimator is on the average unbiased, but this is being unbiased only with respect to the subjective probability. It is biased in any other sense.
2. The Bayesian is required by his paradigm to plug-in the same estimator in estimating all functionals of a non-Euclidean parameter under quadratic loss function. The non-existence of a PIE, a universal estimator that can be plugged-in, makes the Bayesian paradigm too inflexible. This was shown in Section 4.
3. The fact that the Bayes estimator assumes that elements that *a priori* have mean 0, can be considered without harming the final results, played a rule in the failure of the Bayes estimator in the partial linear model of Section 3.
4. On the other side of the previous point, replacing signal buried deeply by the noise with 0, may bias the estimator when the components of the signal can be estimated without bias and accumulated without a bias and bounded overall variance. See the example of Section 5.

5. Conceptually, the strict likelihood principle (cf. Berger and Wolpert (1988)), causes the Bayesian to ignore auxiliary information that may be used to unbiased the estimator. This was in the center of the argument of Section 2, the CODA example (the missing probability), Section 3, the partial linear model (the information about the roughness), and the stopping procedure in the stopping time example of Section 6.

Real life examples are more complex and less traceable than the toy problem we played with in this paper. As a result, it would be harder to understand what are the subtle implications of the assumptions hidden in the prior. It is very hard to build a really complicated prior. The typical researcher would use a prior in which there is a lot of independent component. However, with many independent or semi-independent component laws like LLN and CLT take effect, and thus what was supposed to be a vague prior is concentrated in a small corner of the parameter space. This invariably would impinge of different estimation problems.

Appendix A: An auxiliary result: independent but correlated series

Much of the analysis in this paper is based on presenting counter examples of parameter values on which a given procedure fails. This is satisfactory from a minimax frequentist point of view: one example is enough to argue that the result depends on the unknown parameter and is not uniformly valid, or asymptotic minimax. However, this may not convince the Bayesian, which may claim that the counter example is *a priori* unreasonable. A typical example of the argument was presented in the CODA example of Section 2. It can be characterized by having two *a priori* independent processes (p and g in the example), which happened to be "similar". One may thought that for the Bayesian this is a very unlikely event. After all, he assumes that they are independent, one depends on the biology and the other on the budget constraints. In this section we argue that actually this can be a likely event. Thus, Harmeling and Toussaint (2007) write: "Let us now get to the core of Robins and Ritov [1997]. The authors consider uniform unbiasedness of an estimator. This means that the estimator has to be unbiased for every possible choice of θ and ξ. In the experiment we performed above, though, we chose θ and ξ independently and thus it was very unlikely that we ended up with an accidentally correlated θ and ξ, e.g., where θ tends to be large whenever also ξ is (or inversely)." We claim that this criticism is ignoring the fact that two processes can be independent, while most likely, have high (cross-)correlation. This would be the case if they are not mixing, and have similar autocorrelation function. We elaborate on this in this appendix.

Suppose $U_1, \ldots, U_n$ and $V_1, \ldots, V_n$ are two independent simple random walks. Then of course U_n and V_n are uncorrelated. But we may consider the correlation between these two series $R = n^{-1} \sum_{i=1}^{n} (U_i - \bar{U}_n)(V_i - \bar{V}_n)$, where $\bar{U}_n$ and $\bar{V}_n$ are the empirical means of the two series respectively. R is a random variable, and clearly it has mean 0. However, it is far from being close to 0 even if n is large. In fact asymptotically it is distributed almost uniformly on most of the

interval $(-1, 1)$. McShane and Wyner (2011) use this argument to argue against some standard analysis of historical data in climate research. The reason for this some what surprising fact is that random walks and Brownian motions are less wild than they are some time pictured. In fact given U_n, the best guess of $U_{\lfloor n/2 \rfloor}$ is $U_n/2$, and the sequence tends to be, very roughly speaking, monotone. But if both $U_1, \ldots, U_n$ and $V_1, \ldots, V_n$ are "somewhat" monotone, then they are serially correlated, maybe positively correlated, maybe negatively so, but usually not uncorrelated.

Consider now two general independent mean 0 random non-mixing sequences $U_1, \ldots, U_n$ and $V_1, \ldots, V_n$ Suppose that the two sequence have some autocorrelation functions $A(i, j)$ and $B(i, j)$. We do not assume that the series are stationary, and we do not know their autocorrelation function. The picture we have in mind is that each (U_i, V_i) is a characteristic of points in a large graph, and neighbor nodes are highly correlated, but we do not know the neighborhood structure of the graph. Let

$$R = \text{x-cov}(U_\cdot, V_\cdot) \equiv \frac{1}{n}\sum_{i=1}^{n} U_i V_i - \frac{1}{n^2}\sum_{i=1}^{n} U_i \sum_{i=1}^{n} V_i,$$

where x-cov stands for Cross COVariance. Then $ER = 0$, while direct calculations give:

$$\begin{aligned}\text{Var}(R) &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} A(i,j)B(i,j) \\ &= \frac{1}{n}\sum_{i=1}^{n} \text{x-cov}(A(i,\cdot), B(i,\cdot) \mid i) - \text{x-cov}(\frac{1}{n}\sum_{j=1}^{n} A(\cdot, j), \frac{1}{n}\sum_{j=1}^{n} B(\cdot, j)).\end{aligned}$$

To get an impression suppose that $n^{-1}\sum_{j=1}^{n} A(i,j) \equiv n^{-1}\sum_{j=1}^{n} B(i,j) \equiv c$. Then we get

$$\text{Var}(R) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\big(A(i,j) - c\big)\big(B(i,j) - c\big)$$

Clearly if the two series are mixing, and $\sum_j A(i,j) = \sum_j B(i,j) = \mathcal{O}(1)$ then $\text{Var}(R) = \mathcal{O}(n^{-1})$. However, if they are not mixing, and have similar autocorrelation functions, then most realization of these two series would be serially correlated.

Appendix B: Bayesian estimator for the CODA model when the weight function is known

Let

$$G_j = \left\{ i : g_i = g(x_i) \in (jmn^{-1/2}, (j+1)mn^{-1/2}) \right\},$$

for some constant m. The prior is defined so that h_i is constant on G_j, sampled from a non-informative prior, such that the values $\tilde{h}_j$ on the different sets are independent. In this case:

$$\begin{aligned} n^{-1}E\left(\sum_i h_i \,\middle|\, \text{data}\right) &= n^{-1}E\left(\sum_j \tilde{h}_j|G_j| \,\middle|\, \text{data}\right) \\ &= n^{-1}\sum_j \frac{\sum_{G_j} Y_i R_i}{\sum_{G_j} R_i}|G_j|. \end{aligned}$$

But the fraction on the RHS is the proportion of $\sum_{G_j} R_i$ put randomly in either $\{i : Y_i = 1, i \in G_j\}$ or $\{i : Y_i = 0, i \in G_j\}$ which falls in the first of these sets. If g_i was really constant over G_j, this would correspond to a hypergeometric distribution. Since g is not really constant over G_j,we should add an error term and get

$$E\left(\frac{\sum_{G_j} Y_i R_i}{\sum_{G_j} R_i}|G_j| \,\middle|\, \sum_{G_j} Y_i, |G_j|\right) = \frac{\sum_{G_j} Y_i}{|G_j|}|G_j| + \Delta_j = \sum_{G_j} Y_i + \Delta_j,$$

where $|\Delta_j| < m|G_j|n^{-1/2}$. Since the cells are independent, the Bayesian estimator is $\sqrt{n}$-consistent.

Appendix C: Proofs

Proof of Proposition 4.1. Clearly

$$\begin{aligned} \mathrm{E}\sum_{i=1}^{\infty}(\tilde{\beta}_i - \beta_i)^2 &= \lfloor n^{1/2\alpha} \rfloor / n + \sum_{i > n^{1/2\alpha}} \beta^2 \\ &\leq n^{-(2\alpha-1)/2\alpha} + \sum_{i > n^{1/2\alpha}} i^{-2\alpha} \\ &\leq \frac{2\alpha}{2\alpha - 1} n^{-(2\alpha-1)/2\alpha}. \end{aligned}$$

That this is the minimax risk is established by considering the Bayes prior which makes $\beta_1, \beta_2, \ldots$ independent, $P(\beta_i = \pm i^{-\alpha}) = 1/2$ □

Proof of Lemma 4.2. First note that because of the monotone likelihood ratio property, $\hat{\vartheta}(x)$ is a monotone increasing function of x.

$$\begin{aligned} 1 + \dot{b}_t &= \frac{\partial}{\partial \vartheta} \mathrm{E}_\vartheta \, \mathrm{E}_\pi(\Theta|X) \\ &= \frac{\partial}{\partial \vartheta} \mathrm{E}_\vartheta \, \frac{\int t e^{-(X-t)^2/2\sigma^2} dt}{\int t e^{-(X-t)^2/2\sigma^2} dt} \end{aligned}$$

where E_ϑ is the expectation assuming the true value of the parameter is ϑ, and (Θ, X) is a pair of random variables such that $\Theta \sim \pi$, and $X|\Theta \sim N(\Theta, \sigma^2)$, and E_π is the expectation under their joint distribution. Note that E_π is a formal expression, since we assume that $X \sim N(\vartheta, \sigma^2)$. Let $Z \sim N(0, \sigma^2)$ then

$$\begin{aligned}
1 + \dot{b}_t &= \frac{\partial}{\partial\vartheta}\mathrm{E}\,\frac{\int t e^{-(Z+\vartheta-t)^2/2\sigma^2}d\pi(t)}{\int e^{-(Z+\vartheta-t)^2/2\sigma^2}d\pi(t)} \\
&= \frac{1}{\sigma^2}\mathrm{E}\Big\{\frac{\int t(t-Z-\vartheta)e^{-(Z+\vartheta-t)^2/2\sigma^2}d\pi(t)}{\int e^{-(Z+\vartheta-t)^2/2\sigma^2}d\pi(t)} \\
&\qquad - \frac{\int t e^{-(Z+\vartheta-t)^2/2\sigma^2}d\pi(t)}{\int e^{-(Z+\vartheta-t)^2/2\sigma^2}d\pi(t)}\frac{\int (t-Z-\vartheta)e^{-(Z+\vartheta-t)^2/2\sigma^2}d\pi(t)}{\int e^{-(Z+\vartheta-t)^2/2\sigma^2}d\pi(t)}\Big\} \\
&= \frac{1}{\sigma^2}\mathrm{E}_\vartheta\{\mathrm{Var}(\Theta \mid X)\},
\end{aligned}$$

Hence $0 \le 1 + \dot{b}_\vartheta \le (a/\sigma)^2$, or $\dot{b}_\vartheta \in [-1, -(1-(a/\sigma)^2)]$. The lemma follows the mean value theorem.

□

Proof of Theorem 7.1. The proof is based on the two lemmas that follows. Suppose the posterior is not d_n consistent on $\mathcal{H}'$ with $\pi(\mathcal{H}') > 0$. Then by Lemma C.1, (2) must hold for $\eta_0 \in \mathcal{H}'$. By Lemma C.2, (4) must hold. But (4) contradict that $\pi(\mathcal{H}) = 1$, since then from all M: $\pi\{\eta : P_\eta(\sqrt{\eta} - \tilde{\eta}_n geM)\} > 0$.

□

Lemma C.1 *Suppose*

1. *There is a statistic $\tilde{\eta}_n$ such that $\limsup_n P_{\eta_0}(d_n(\tilde{\eta}_n, \eta_0) \ge M) \to 0$ as $M \to \infty$.*
2. *For all $M < \infty$: $\limsup_n P_{\eta_0}\big(\pi_{X_n}(d_n(\eta_n, \eta_0) \ge 2M) \ge 2\varepsilon\big) \ge 2\delta$.*

Then there is M which may depend on η_0 such that

$$\limsup_n P_{\eta_0}\big(\pi_{X_n}(d_n(\eta_n, \tilde{\eta}_n) \ge M) \ge \varepsilon\big) \ge \delta. \tag{2}$$

Proof.

$$\begin{aligned}
&P_{\eta_0}\big(\pi_{X_n}(d_n(\eta_n, \tilde{\eta}_n) \ge M) \ge \varepsilon\big) \\
&\quad \ge P_{\eta_0}\big(\{\pi_{X_n}(d_n(\eta_n, \eta_0) \ge 2M) \ge 2\varepsilon\} \cap \{d_n(\tilde{\eta}_n, \eta_0) \le M\}\big) \\
&\quad \ge P_{\eta_0}\big(\pi_{X_n}(d_n(\eta_n, \eta_0) \ge 2M) \ge 2\varepsilon\big) - P_{\eta_0}\big(d_n(\tilde{\eta}_n, \eta_0) \ge M\big)
\end{aligned}$$

By assumption the lim-sup of the first term on the RHS is bounded by 2δ, while we can choose M large enough such that the second term is bounded by δ for all large enough n. The lemma follows. □

Lemma C.2 *Suppose there exist a statistic $\tilde{\eta}_n$, M, $\varepsilon, \delta > 0$ such that*

$$P_{\eta_0}\big(\pi_{X_n}(d_n(\tilde{\eta}_n, \eta_n) \ge M) \ge \varepsilon\big) \ge \delta \tag{3}$$

for all $\eta_0 \in \mathcal{H}'$ *and* $\pi(\mathcal{H}') \geq \gamma > 0$. *Then for all* $M < \infty$*:*

$$P(d_n(\tilde{\eta}_n, \eta_0) \geq M) \geq \varepsilon\delta\gamma, \tag{4}$$

Proof. Obviously, if U, V, W are three random variables, then $E(E(E(U|V)|W) = E(U)$. Hence taking the expectation of (3), we obtain (4). □

References

Bayarri, M. and Berger, J. (2004). The interplay between Bayesian and frequentist analysis. *Statistical Science*, **19**, 58–80.

Berger, J. (2006a). The case for objective Bayesian analysis. *Bayesian Analysis*, **1**, 385–402.

Berger, J. (2006b). Rejoinder. *Bayesian Analysis*, **1**, 457–464.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis* (2nd ed.). Springer-Verlag, New York.

Berger, J. O. and Wolpert, R. L. (1988). *The Likelihood Principle: A Review, Generalizations, and Statistical Implications* (2nd ed.)., volume 6 of *Lecture Notes—Monograph Series*. IMS, Hayward, California.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. (1998). *Efficient and adaptive estimation in semiparametric models*. Springer-Verlag, New York.

Bickel, P. J. and Kleijn, B. J. (2012). The semiparametric Bernstein-von Mises theorem. *Ann. Statist.*, **2012**, To appear.

Bickel, P. J. and Ritov, Y. (1989). Estimation of squared integrated density derivatives. *Sankhya (1989).*, **A50**, 381–393.

Bickel, P. J. and Ritov, Y. (2003). Nonparametric estimators which can be "plugged-in". *Ann. Statist.*, *31*(4), 1033–1053.

Cochran, W. G. (1977). *Sampling Techniques* (3rd ed.). Wiley, New York.

Cox, D. (1993). An analysis of Bayesian inference for nonparametric regression. *Ann. Statist.*, **21**, 903–924.

Diaconis, P. and Freedman, D. (1998). Consistency of Bayes estimates for nonparameteric regression: Normal theory. *Bernoulli*, **4**, 411–444.

Engle, R. F., Granger, C. W. J., Rice, J., and Weiss, A. (1986). Nonparametric estimates of the relation between weather and electricity sales. *J. Amer. Statist. Assoc.*, **81**, 310–320.

Freedman, D. (1993). On the asymptotic behavior of Bayes estimates in the discrete case i. *Ann. Math. Statist.*, **34**, 1386–1403.

Freedman, D. (1999). On the Bernstein-von Mises theorem with in nite dimensional parameters. *Ann. Statist.*, **27**, 1119–1140.

Ghosal, S., Ghosh, J., and van der Vaart, A. (2000). Convergence rates of posterior distributions. *Ann. Statist.*, **28**, 500–531.

Goldstein, M. (2006). Subjective Bayesian analysis: Principles and practice. *Bayesian Analysis*, **1**, 403–420.

Harmeling, S. and Toussaint, M. (2007). Bayesian estimators for robins-ritov's

problem. Technical report, University of Edinburgh, School of Informatics Research Report EDI-INF-RR-1189.

Kleijn, B. and van der Vaart, A. (2006). Misspecification in infinite-dimensional Bayesian statistics. *Ann. Statist.*, **34**, 837–877.

Le Cam, L. and Yang, G. (1990). *Asymptotics in Statistics: Some Basic Concepts*. Springer, New York.

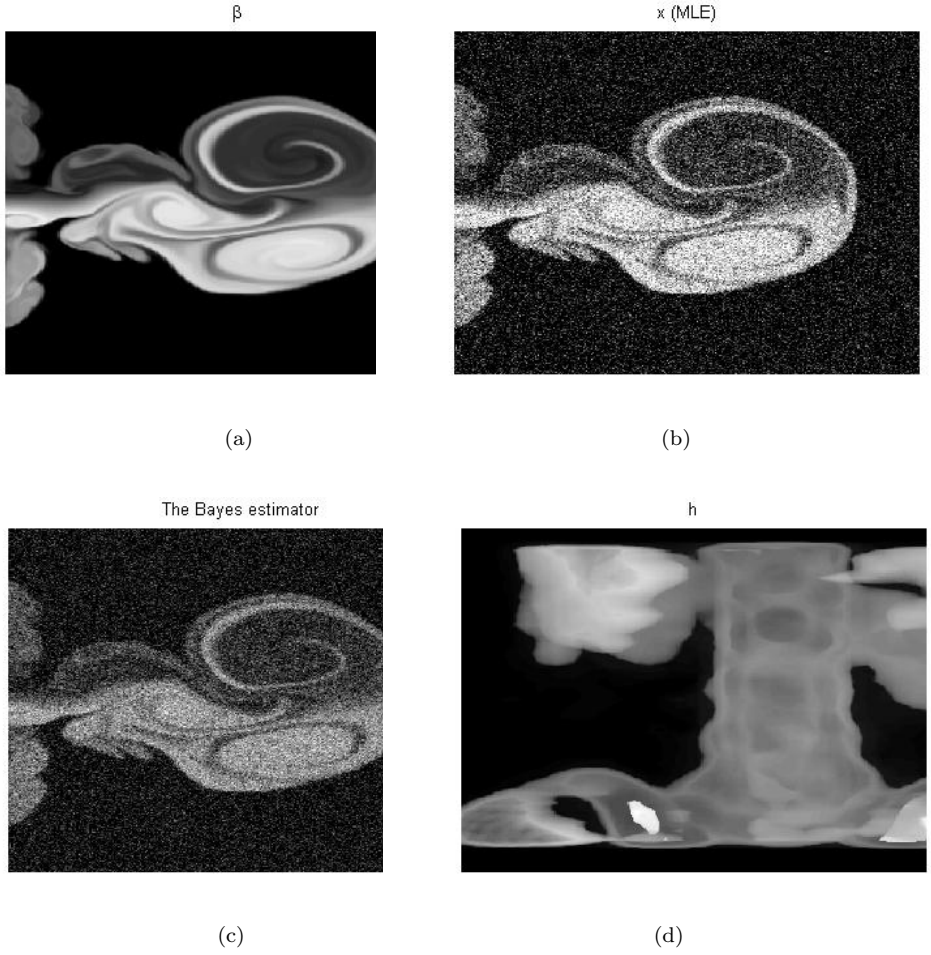
Lehmann, E. and Casella, G. (1998). *Theory of Point Estimation*. Springer, New York.

McShane, B. B. and Wyner, A. J. (2011). A statistical analysis of multiple temperature proxies: Are reconstructions of surface temperatures over the last 1000 years reliable? *Ann. Appl. Stat.*, **5**, 5–44.

Robins, J. M. and Ritov, Y. (1997). Toward a curse of dimensionality appropriate (coda) asymptotic theory for semiparametric models. *Statistics in Medicine*, **17**, 285–319.

Schick, A. (1986). On efficient estimation in regression models. *Ann. Statist.*, **14**, 1486–1521.

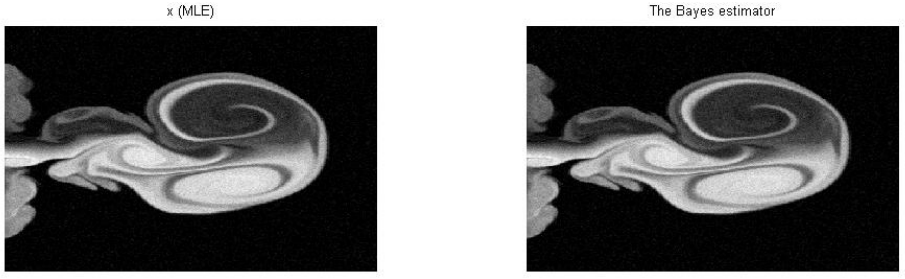
Wasserman, L. (1998). Asymptotic properties of nonparametric Bayesian procedures. In *Lecture Notes in Statistics, vol. 133: Practical Nonparametric and Semiparametric Bayesian Statistics, editors: D. Dey, P. Müller and D. Sinha* (pp. 293–304). Springer, New York.



(a) (b)

(c) (d)

Fig 1: Computing linear functionals. (a) The vector $\boldsymbol{\beta}$; (b) The vector $\boldsymbol{\beta} + \boldsymbol{\varepsilon}$; (c) The Bayes estimator; (d) The functional h.



(a) (b)

Fig 2: (a) $\boldsymbol{\beta} + \boldsymbol{\varepsilon}$; (b) The Bayes estimator.